**Supplementary Materials** to:

Authors (Year). Hebb repetition learning in adolescents with intellectual disabilities. *Research in Developmental Disabilities*.

This supplement contains the following four sections:

**Section 1:** Detailed results examining the reliability of measuring inter-individual differences in Hebb repetition learning. The key data here are presented in bold.

**Section 2**: Further descriptive results in relation to Figure 2 in the paper.

**Section 3:**  The complete results tables of the Generalized Linear Mixed Effects models described in the paper, including the full random part of the models. Coefficients of the random part are not displayed in the paper, as all research questions were focussed on the fixed part. During model estimation, the random effects were included to properly reflect the hierarchical structure of the data (trials were nested within participants) and to properly model individual differences, as long as the available data allowed that random effects could be modelled.

**Section 4:**  Models with normalized Levenshtein Distance (an alternative scoring method for Hebb repetition learning tasks) as the dependent variable.

# **Section 1: Examining the reliability of measuring** **inter-individual differences in Hebb repetition learning**

Reliability here represents how consistently inter-individual differences in Hebb repetition learning (HRL) can be measured. In mixed effect models the residuals of the random effects reflect how an individual differs from the group mean (fixed effects). To derive reliability estimates for HRL, we, therefore, estimated separate GLMMs for session 1 and session 2, extracted the residuals of the random effects for the list type x position interaction (indicating HRL) from each model and computed retest-reliabilities.

In the preregistration document we specified how we would examine reliability (please see hypothesis 3; <https://osf.io/gkpwh/>). In the current article, research questions 1 to 3 correspond to hypotheses 1a to 1c of the preregistration (note that due to word limits we will report the analyses preregistered for hypothesis 2 in a separate paper):

“For Hypothesis 3, we will assess split-half reliability and test-retest reliability between Hebb-or-Filler List X Trial Position residuals of sessions 1 and Hebb-or-Filler List X Trial Position residuals of session 2 using correlations. Our measure of reliability will be derived from entering the correlation of half1 with half2 into the Spearman-Brown-Formula for a doubling in test length.   
Data sets include: 1) only session 1, but both groups, both types of material   
2) only session 2, but both groups, both types of material   
Fixed effects: Hebb-or-Filler List Trial Position Material Hebb-or-Filler List X Trial Position   
Random effects: a = Hebb-or-Filler List X Trial Position;” “b = random intercepts, c = random effect for material, d = random effect for trial position”, “b – d if model fit improved (see above)”

***Task reliability for research questions 1 and 2***

Please note that this analysis plan adequately estimates reliability for research question 1 (and 2) in the current paper as the model (fixed effects and random effects) and the data structure (both trials from verbal and visuospatial material) are the split-half version of the models addressing these research questions (compare paper Table 2, supplement Table S1). The models used to analyse the research questions and to estimate reliability contain: random intercepts (b); random effects for material (c); random effects for position (d); and a random effect for the list type x trial position interaction (a). Therefore, they model inter-individual differences in the general performance level (b), in recall level differences between verbal and visuospatial material (c), and in recall development in the task averaged across Hebb and filler trials (d), in addition to inter-individual differences in recall improvements in Hebb trials (in contrast to filler trials) (a). Consequently, taking the residuals (i.e., the individual deviations from the fixed effect representing the group mean) of the random effect variance of the list x position interaction should be the appropriate indicator for individual differences in HRL, controlling for other individual differences in recall performance.

Table S1. GLMMs on recall performance across trial positions for Hebb vs. filler lists to address ***reliability across verbal and visuospatial material*** for research question 1 (therefore same fixed effects and random effects structure as Table S4).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Session 1** | | |  | **Session 2** | | |
| Fixed Part | β | 95% CI | *p* |  | β | 95% CI | *p* |
| Intercept 1 (no vs. item recall) | 1.15 | [1.03, 1.26] | < .001 |  | 1.15 | [0.96, 1.32] | < .001 |
| Intercept 2 (item vs. position recall) | -0.74 | [-0.91, -0.57] | < .001 |  | -0.57 | [-0.78, -0.37] | < .001 |
| Material (verbal = +1 vs. visuospatial = -1) | 0.48 | [0.42, 0.54] | < .001 |  | 0.35 | [0.26, 0.43] | < .001 |
| List (Hebb = +1 vs. Filler = -1) | 0.06 | [0.00, 0.11] | .06 |  | 0.14 | [0.08, 0.21] | < .001 |
| Trial Position (1st = 0, 2nd = 1 …) | 0.01 | [-0.01, 0.02] | .55 |  | -0.02 | [-0.04, 0.00] | .13 |
| List x Position (Hebb effect) | 0.09 | [0.07, 0.11] | < .001 |  | 0.06 | [0.04, 0.09] | < .001 |
| Random Part | u | 95% CI |  |  | u | 95% CI |  |
| *Random effect variances* |  |  |  |  |  |  |  |
| Intercept 1 | 0.005 | [0.003, 0.007] |  |  | 0.008 | [0.005, 0.011] |  |
| Intercept 2 | 0.004 | [0.002, 0.007] |  | 0.005 | [0.002, 0.008] |
| Material | 0.681 | [0.485, 0.931] |  |  | 0.904 | [0.642, 1.247] |  |
| Trial Position | 0.270 | [0.182, 0.390] |  |  | 0.628 | [0.433, 0.89] |  |
| List x Position | 0.064 | [0.042, 0.095] |  |  | 0.133 | [0.091, 0.192] |  |
| *Covariances* |  |  |  |  |  |  |  |
| Intercept 1 – Intercept 2 | 0.004 | [0.002, 0.006] |  |  | 0.002 | [0.000, 0.004] |  |
| Intercept 1 – Material | 0.044 | [0.029, 0.063] |  |  | 0.063 | [0.041, 0.090] |  |
| Intercept 1 – Position | 0.027 | [0.017, 0.041] |  |  | 0.051 | [0.033, 0.075] |  |
| Intercept 1 – List x Position | 0.000 | [-0.005, 0.006] |  |  | -0.002 | [-0.01, 0.007] |  |
| Intercept 2 – Material | 0.034 | [0.019, 0.052] |  |  | -0.002 | [-0.026, 0.018] |  |
| Intercept 2 – Position | 0.019 | [0.009, 0.029] |  |  | -0.008 | [-0.030, 0.009] |  |
| Intercept 2 – List x Position | -0.001 | [-0.006, 0.005] |  |  | 0.005 | [-0.004, 0.014] |  |
| Material – Position | 0.365 | [0.244, 0.519] |  |  | 0.653 | [0.442, 0.933] |  |
| Material – List x Position | -0.032 | [-0.092, 0.019] |  |  | -0.062 | [-0.158, 0.025] |  |
| Position – List x Position | -0.031 | [-0.072, 0.007] |  |  | -0.069 | [-0.151, 0.004] |  |

Note. Sample sizes for session 1: *N*Participants = 94, *N*Trials = 3008, *N*Items= 13952; for session 2: *N*Participants = 86, *N*Trials = 2736, *N*Items = 12752

From the two models presented in Table S1 the residuals of the List x Position variances were extracted as the HRL indicator in session 1 and as a retest as HRL indicator in session 2.

**The retest correlation was: *r*tt = .47, 95% CI [.29, .62], p < .001**. **According to the Spearman-Brown formula the split-half-reliability for assessing Hebb learning based on session 1 and 2 therefore was .64. for the analyses on research questions 1 and 2.**

***Task reliability for research question 3***

Turning to research question 3 in the current paper, the reliability estimate derived from the two models on both materials (see Table S1) is not the most appropriate estimate for separately analysing HRL with verbal vs. visuospatial material (to answer research question 3). Models differ in their fixed and random effects and the pool of data they analyse (compare paper Table 3, supplement Table S2). Reliability might be lower, because interindividual differences in Hebb learning have to be estimated from half the number of trials (either verbal or visuospatial trials). Reliability might be higher because Hebb learning is estimated from more homogeneous trials (either verbal or visuospatial trials). Reliability might be similar if both potential effects cancel out. Therefore, equivalent split-half reliability analyses were performed for verbal material and for visuospatial material even though these analyses had not been preregistered.

Models for research question 3 cannot include a random effect for material, because the models are either based on verbal or on visuospatial material. The models on verbal material include random effects for the list x position interaction, the intercepts and position (compare Tables S5 and S2). The models for visuospatial material do not support including both random effects for position and for the list x position interaction; model estimations fail. As visuospatial memory lists comprise only three or four items compared to five of six items in verbal memory lists we might have collected too few data points per child to consider all relevant interindividual differences. Model estimations succeed, however, when models contain only the random effect for the list x position in addition to the intercepts (compare Tables S5 and S3). However, as no random effect for position effect could be included the question remained whether the residuals of the list x position interaction should be interpreted as indicating differences in Hebb learning, differences in a general position effect or a combination of both effects (but with too few data points per child to disentangle both effects).

To decide between the three interpretations, one approach is to separate performance gains on Hebb trials from potential performance declines in filler trials due to interference – this can deliver helpful insights (compare post-hoc analyses, paper manuscript p. 24). At the group level, the fixed effects showed that for verbal material both recall gains on Hebb trials and recall declines on filler trials were present. For verbal material, models supported random effect variances for position effects both on Hebb trials and on filler trials and residuals from session 1 and 2 were correlated (Hebb gains: *r*tt = .62, 95% CI [.47, .73], *p* < .001; filler interference: *r*tt = .56, 95% CI [.39, .69], *p* < .001). In contrast, for visuospatial material fixed effects indicated that at the group level only performance gains on Hebb trials were present, but performance on filler trials did not change. These models typically supported including a random position effect for Hebb trials[[1]](#footnote-1), but no model supported a random position effect for filler trials. Not finding a position effect for visuospatial filler trials (neither in the fixed nor the random part) indicates that for visuospatial material one random effect for capturing interindividual differences in recall development across trial positions within the Hebb-and-Filler-task is sufficient to properly model the data and that this effect captures interindividual differences in Hebb learning. Additionally, having fewer items to be recalled seems to make assessing interindividual differences in Hebb learning less robust for our visuospatial material compared to our verbal material.

Table S2. GLMMs on recall performance on **verbal material** across trial positions for Hebb vs. filler lists to address **reliabilityfor research question 3** (therefore same fixed effects and random effects structure as Table S5 – verbal material).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Session 1** | | |  | **Session 2** | | |
| Fixed Part | β | 95% CI | *p* |  | β | 95% CI | *p* |
| Intercept 1 (no vs. item recall) | 1.91 | [1.74, 2.07] | < .001 |  | 1.68 | [1.45, 1.89] | < .001 |
| Intercept 2 (item vs. position recall) | -0.27 | [-0.47, -0.08] | .007 |  | -0.21 | [-0.49, 0.03] | .10 |
| List (Hebb = +1 vs. Filler = -1) | 0.06 | [-0.02, 0.15] | .12 |  | 0.13 | [0.05, 0.22] | .002 |
| Trial Position (1st = 0, 2nd = 1 …) | 0.00 | [-0.03, 0.03] | .78 |  | -0.03 | [-0.06, 0.01] | .10 |
| List x Position (Hebb effect) | 0.14 | [0.11, 0.17] | < .001 |  | 0.09 | [0.06, 0.13] | < .001 |
| Random Part | u | 95% CI |  |  | u | 95% CI |  |
| *Random effect variances* |  |  |  |  |  |  |  |
| Intercept 1 | 0.015 | [0.010, 0.021] |  |  | 0.014 | [0.009, 0.021] |  |
| Intercept 2 | 0.012 | [0.007, 0.020] |  | 0.016 | [0.009, 0.025] |
| Position | 0.809 | [0.559, 1.139] |  |  | 1.216 | [0.844, 1.722] |  |
| List x Position | 0.463 | [0.295, 0.701] |  |  | 0.844 | [0.561, 1.224] |  |
| *Covariances* |  |  |  |  |  |  |  |
| Intercept 1 – Intercept 2 | 0.010 | [0.006, 0.016] |  |  | 0.004 | [0.000, 0.010] |  |
| Intercept 1 – Position | 0.066 | [0.038, 0.100] |  |  | 0.079 | [0.045, 0.122] |  |
| Intercept 1 – List x Position | 0.046 | [0.023, 0.074] |  |  | 0.065 | [0.036, 0.102] |  |
| Intercept 2 – Position | 0.040 | [0.011, 0.071] |  |  | -0.030 | [-0.078, 0.009] |  |
| Intercept 2 – List x Position | 0.006 | [-0.020, 0.030] |  |  | -0.035 | [-0.078, 0.000] |  |
| Position – List x Position | 0.447 | [0.272, 0.689] |  |  | 0.870 | [0.574, 1.278] |  |

Note. Sample sizes for session 1: *N*Participants = 94, *N*Trials = 1504, *N*Items= 8224; for session 2: *N*Participants = 86, *N*Trials = 1376, *N*Items = 7568

Table S3. GLMMs on recall performance on **visuospatial material** across trial positions for Hebb vs. filler lists to address **reliabilityfor research question 3** (therefore same fixed effects and random effects structure as Table S5 – visuospatial material).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Session 1** | | |  | **Session 2** | | |
| Fixed Part | β | 95% CI | *p* |  | β | 95% CI | *p* |
| Intercept 1 (no vs. item recall) | 0.45 | [0.29, 0.62] | < .001 |  | 0.6 | [0.40, 0.80] | < .001 |
| Intercept 2 (item vs. position recall) | -1.14 | [-1.36, -0.92] | < .001 |  | -0.91 | [-1.14, -0.68] | < .001 |
| List (Hebb = +1 vs. Filler = -1) | 0.04 | [-0.05, 0.13] | 0.41 |  | 0.16 | [0.06, 0.25] | .002 |
| Trial Position (1st = 0, 2nd = 1 …) | 0.03 | [0.01, 0.05] | .004 |  | 0.01 | [-0.01, 0.04] | .22 |
| List x Position (Hebb effect) | 0.04 | [0.02, 0.07] | < .001 |  | 0.04 | [0.01, 0.06] | .02 |
| Random Part | u | 95% CI |  |  | u | 95% CI |  |
| *Random effect variances* |  |  |  |  |  |  |  |
| Intercept 1 | 0.003 | [0.002, 0.005] |  |  | 0.007 | [0.004, 0.012] |  |
| Intercept 2 | 0.457 | [0.315, 0.643] |  | 0.697 | [0.480, 1.000] |
| List x Position | 0.866 | [0.623, 1.178] |  |  | 1.068 | [0.758, 1.497] |  |
| *Covariances* |  |  |  |  |  |  |  |
| Intercept 1 – Intercept 2 | 0.033 | [0.021, 0.050] |  |  | 0.067 | [0.044, 0.099] |  |
| Intercept 1 – List x Position | 0.044 | [0.028, 0.065] |  |  | 0.080 | [0.053, 0.116] |  |
| Intercept 2 – List x Position | 0.579 | [0.408, 0.801] |  |  | 0.777 | [0.536, 1.123] |  |

Note. Sample sizes for session 1: *N*Participants = 94, *N*Trials = 1504, *N*Items= 5728; for session 2: *N*Participants = 85, *N*Trials = 1360, *N*Items = 5184

From the four models presented in Table S2 and S3 the residuals of the List x Position variances were extracted as the HRL indicator in session 1 and at retest as the HRL indicator in session 2, for both materials separately.

**The retest correlations were *r*tt = .63, 95% CI [.48, .74], *p* < .001 for verbal material and *r*tt = .69, 95% CI [.56, .79], *p* < .001 for visuospatial material. According to the Spearman-Brown formula the split-half-reliabilities for assessing Hebb learning separately for each material but based on session 1 and 2 therefore were .77 (verbal) and .82 (visuospatial) for the analyses on research question 3. The four correlations between Hebb learning assessed with verbal and with visuospatial material ranged from *r* = .45, 95% CI [.26, .60] to *r* = .59, 95% CI [.44, .71].**

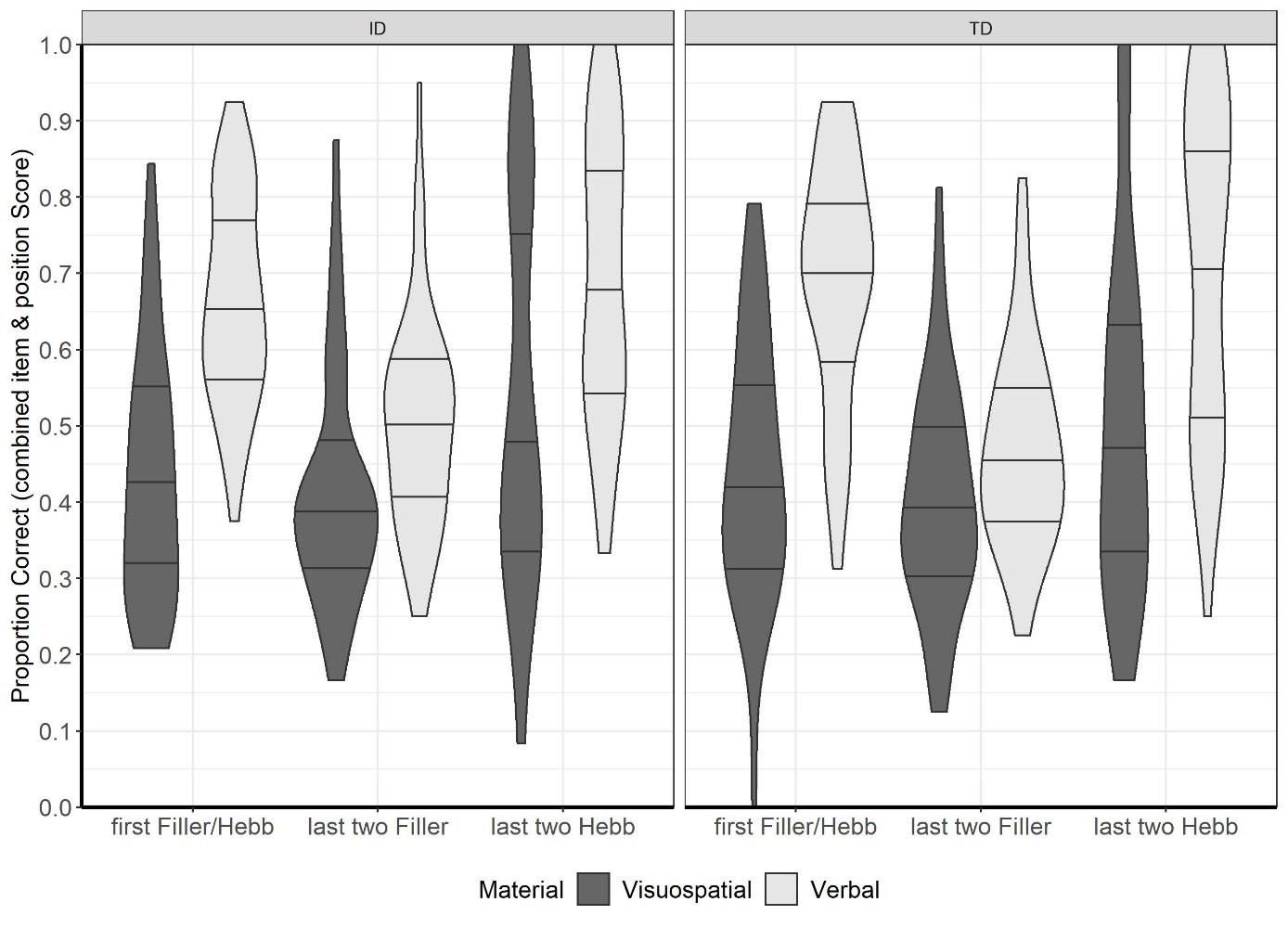
**Section 2: Further descriptive results in relation to Figure 2 in the paper**

Figure 2 of the paper indicated that participants started into the Hebb experiment with an average recall performance of about .66 to .70 with verbal material and about .43 on visuospatial material. Given that some participants received shorter lists (visuospatial list length 3 items, *N*ID = 9, *N*TD = 9; verbal list length 5 items, *N*ID = 24, *N*TD = 26) and others received longer lists (visuospatial list length 4 items; *N*ID = 38, *N*TD = 38; verbal list length 6 items; *N*ID = 23, *N*TD = 21), and given that participants were awarded partial points when they remembered an item (in any position) and additional points when they item was remembered in correct position, the recall scores can be hard to interpret. Here, we provide more information to facilitate interpretation and to explore whether issues with ceiling or floor effects were present in the HRL tasks.

Was there room for improvement for the average participant? The score of .66 in the verbal task translates into a participant with longer lists of six nameable pictures scoring 8 out of 12 partial recall points, e.g., by remembering three items in correct position (two points each) and two further items in any position (1 point each). The score of .70 in the verbal task translates into a participant with shorter lists of five nameable pictures scoring 7 out of 10 partial recall points, e.g., by remembering three items in correct position (two points each) and a further item in any position (1 point each). The average recall score of .43 in the visuospatial task is equivalent to a participant with longer list of four nonsense pictures scoring 3 out of 8 partial recall points in session 1 and 4 out of 8 partial recall points in session 2 and equivalent to a participant with shorter list (only 19% of participants received three item lists) scoring 3 out of 6 and 2 out of 6 partial recall points. Therefore, for the average participant there was both room for improving recall performance (e.g., by HRL) or for declining performance (e.g., by interference or by fatigue).

To explore whether some participants started the verbal task at ceiling with perfect recall in correct positions or some participants started the visuospatial at the floor with no recall at all, distributions of scores are displayed in Figure S1. To avoid very high or very low performance levels skewing the number of extreme values for each participant by chance, performance on (typically four) trials was averaged. For the starting performance, these were the first filler and the first Hebb trial in session 1 and 2. As both were the first trials with a new set of items, they were logically equivalent (and performance was empirically comparable). For performance at the end of the Hebb experiment, the 7th and 8th filler trials from both sessions were averaged, as were the 7th and 8th Hebb trials from both sessions.

Figure S1. Violin plots (with lines indicating the 25th, 50th and 75th quantile) displaying recall scores at the start and the end of the Hebb experiment

****

As can be seen, no participant started with perfect recall in the verbal condition, but 25% in both groups started with high recall scores of .80 or above. Only one participant started with no recall at all in the visuospatial condition, but 25% in both groups started with low recall score of about .30 or below. At the end of the experiment, 8 participants (= 9%) achieved perfect recall on Hebb trials in the verbal task and 4 participants (= 4%) did so in the visuospatial task. The elongated distributions of recall on the last Hebb trials are a first indication, that some participants benefitted from the repetitions of Hebb trials and improved recall, while others did not. In general, the figure demonstrates that there was room for performance changes in either direction for nearly all participants.

Please note that we did not analyse the displayed proportion corrects sores with ANOVAs or with Linear Mixed Models because the non-normal distributions of the scores and scores below 0.00 and above 1.00 being impossible would violate the assumptions of those analyses. We therefore analysed the data with Generalized Linear Mixed Models for ordinal data, modelling the likelihood that participants recall an item at all (in any position) and recall an item in position. For these models there is no inherent problem if for certain participants under a certain combination of predictors (e.g., List type = Hebb & high trial position) (near) perfect recall is predicted.

# **Section 3: The complete results tables of the Generalized Linear Mixed Effects models described in the paper**

Table S4. GLMMs on recall performance across trial positions for Hebb vs. filler lists to address research question 1 for the ID and TD groups, additionally displaying the random part of the model not included in the corresponding Table 2 of the paper.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | ID | | |  | TD | | |
| Fixed Part | β | 95% CI | *p* |  | β | 95% CI | p |
| Intercept 1 (no vs. item recall) | 1.08 | [0.89, 1.26] | < .001 |  | 1.15 | [0.96, 1.33] | < .001 |
| Intercept 2 (item vs. position recall) | -0.67 | [-0.93, -0.42] | < .001 |  | -0.67 | [-0.92, -0.42] | < .001 |
| Material (verbal = +1 vs. visuospatial = -1) | 0.41 | [0.33, 0.48] | < .001 |  | 0.42 | [0.32, 0.51] | < .001 |
| List (Hebb = +1 vs. Filler = -1) | 0.16 | [0.10, 0.22] | < .001 |  | 0.03 | [-0.03, 0.09] | .30 |
| Trial Position (1st = 0, 2nd = 1 …) | 0.00 | [-0.02, 0.02] | .71 |  | -0.01 | [-0.04, 0.01] | .29 |
| List x Position (Hebb effect) | 0.05 | [0.03, 0.08] | < .001 |  | 0.09 | [0.07, 0.11] | < .001 |
| Random Part | u | 95% CI | Δ DIC |  | u | 95% CI | Δ DIC |
| *Random effect variances* |  |  |  |  |  |  |  |
| Intercept 1 | 0.004 | [0.003, 0.007] | 1993.6 |  | 0.004 | [0.003, 0.007] | 2100.3 |
| Intercept 2 | 0.704 | [0.459, 1.079] |  | 0.810 | [0.520, 1.250] |
| Material | 0.054 | [0.032, 0.088] | 121.7 |  | 0.105 | [0.064, 0.166] | 248.6 |
| Trial Position | 0.369 | [0.229, 0.570] | 7.9 |  | 0.392 | [0.245, 0.624] | 25.1 |
| List x Position | 0.002 | [0.001, 0.004] | 193.5 |  | 0.004 | [0.002, 0.008] | 183.6 |
| *Covariances* |  |  |  |  |  |  |  |
| Intercept 1 – Intercept 2 | -0.002 | [-0.008, 0.003] |  |  | -0.001 | [-0.009, 0.006] |  |
| Intercept 1 – Material | 0.033 | [0.019, 0.053] |  |  | 0.031 | [0.018, 0.051] |  |
| Intercept 1 – Position | 0.043 | [0.026, 0.070] |  |  | 0.049 | [0.030, 0.078] |  |
| Intercept 1 – List x Position | 0.002 | [0.001, 0.004] |  |  | 0.003 | [0.001, 0.005] |  |
| Intercept 2 – Material | -0.04 | [-0.095, 0.006] |  |  | -0.054 | [-0.134, 0.010] |  |
| Intercept 2 – Position | -0.041 | [-0.113, 0.023] |  |  | -0.059 | [-0.165, 0.030] |  |
| Intercept 2 – List x Position | -0.001 | [-0.006, 0.003] |  |  | 0.004 | [-0.004, 0.012] |  |
| Material – Position | 0.441 | [0.269, 0.693] |  |  | 0.467 | [0.278, 0.758] |  |
| Material – List x Position | 0.011 | [0.000, 0.023] |  |  | 0.007 | [-0.009, 0.023] |  |
| Position – List x Position | 0.016 | [0.003, 0.033] |  |  | 0.019 | [-0.004, 0.044] |  |

Note. Sample sizes for the ID group: *N*Participants = 47, *N*Trials = 2944, *N*Items= 13680; for the TD group: *N*Participants = 47, *N*Trials = 2800, *N*Items = 13024

Δ *DIC* =Difference between the DIC (Deviance Information Criterion) of the model not including the random effect of interest minus the DIC of the model including this random effect; e.g. a model assuming no interindividual differences vs. a model only including random separation parameter; the model only including random separation parameters vs. the model additionally including a random effect for material etc. If Δ *DIC* is small (below 5) the model without the random effects fits the data approximately as well as the model with the effect (see Zhang et al., 2016). Therefore, including each set of parameters improved model fit.

No random effect for list was included. This effect would represent inter-individual differences in recall of Hebb and filler list at the first occurrence. As at this position Hebb lists were not yet repeated, there is no meaningful interpretation of such an effect. The significant fixed effect for List in the ID group is an artefact of the model not present in the data (compare no difference in observed data in Figure 2) that was most likely due to some non-linearity in Hebb learning (stronger effect from trial 1-4 than 5-8).

Table S5. GLMMs conducted on recall performance across trial positions for Hebb vs. filler list for verbal and visuospatial stimuli to address hypothesis 3, additionally displaying the random part of the model not included in the corresponding Table 3 of the paper.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Verbal** (with RS Position) | | |  | **Verbal** (without RS Position) | | |  | **Visuospatial** (without RS Position) | | |
| Fixed Part | β | 95% CI | *p* |  | β | 95% CI | *p* |  | β | 95% CI | *p* |
| Intercept 1 (no vs. item recall) | 1.74 | [1.58, 1.9] | < .001 |  | 1.74 | [1.58, 1.90] | < .001 |  | 0.49 | [0.33, 0.65] | < .001 |
| Intercept 2 (item vs. position recall) | -0.27 | [-0.47, -0.08] | .007 |  | -0.26 | [-0.45, -0.06] | 0.01 |  | -1.04 | [-1.24, -0.85] | < .001 |
| List (Hebb = +1 vs. Filler = -1) | 0.10 | [0.05, 0.16] | < .001 |  | 0.11 | [0.05, 0.17] | < .001 |  | 0.09 | [0.02, 0.16] | .007 |
| Trial Position (1st = 0, 2nd = 1 …) | -0.03 | [-0.05, -0.01] | .01 |  | -0.03 | [-0.04, -0.02] | < .001 |  | 0.02 | [0.01, 0.04] | .008 |
| List x Position (Hebb effect) | 0.10 | [0.08, 0.12] | < .001 |  | 0.09 | [0.07, 0.12] | < .001 |  | 0.03 | [0.02, 0.05] | < .001 |
| Random Part | u | 95% CI | Δ DIC |  | u | 95% CI | Δ DIC |  | u | 95% CI | Δ DIC |
| *Random effect variances* |  |  |  |  |  |  |  |  |  |  |  |
| Intercept 1 | 0.008 | [0.005, 0.011] | N.A. 1) |  | 0.007 | [0.005, 0.009] | N.A. 1) |  | 0.003 | [0.002, 0.004] | N.A. 1) |
| Intercept 2 | 0.006 2) | [0.003, 0.010] |  | 0.562 2) | [0.407, 0.774] |  | 0.511 | [0.364, 0.704] |
| Position | 0.863 | [0.625, 1.185] | 44.5 |  |  | - 2) |  |  |  | - 2) |  |
| List x Position | 0.521 2) | [0.365, 0.731] | 352.5 |  | 0.970 2) | [0.711, 1.306] | 329.8 |  | 0.812 | [0.595, 1.096] | 134.2 |
| *Covariances* |  |  |  |  |  |  |  |  |  |  |  |
| Intercept 1 – Intercept 2 | 0.003 | [0.001, 0.006] |  |  | 0.052 | [0.036, 0.073] |  |  | 0.035 | [0.024, 0.049] |  |
| Intercept 1 – Position | 0.058 | [0.038, 0.083] |  |  |  | - 2) |  |  |  | - 2) |  |
| Intercept 1 – List x Position | 0.045 | [0.029, 0.066] |  |  | 0.063 | [0.043, 0.088] |  |  | 0.043 | [0.030, 0.061] |  |
| Intercept 2 – Position | 0.009 | [-0.013, 0.031] |  |  |  | - 2) |  |  |  | - 2) |  |
| Intercept 2 – List x Position | -0.002 | [-0.021, 0.014] |  |  | 0.603 | [0.421, 0.838] |  |  | 0.582 | [0.416, 0.797] |  |
| Position – List x Position | 0.524 | [0.351, 0.749] |  |  |  | - 2) |  |  |  | - 2) |  |

Note. Sample sizes for the verbal condition: *N*Participants = 94, *N*Trials = 2880, *N*Items= 15792; for the visuospatial condition: *N*Participants = 94, *N*Trials = 2864, *N*Items = 10912

Δ *DIC* =Difference between the DIC (Deviance Information Criterion) of the model not including the random effect of interest minus the DIC of the model including this random effect; e.g. a model assuming no interindividual differences vs. a model only including random separation parameter; the model only including random separation parameters vs. the model additionally including a random effect for material etc. If Δ *DIC* is small (below 5) the model without the random effects fits the data approximately as well as the model with the effect (see Zhang et al., 2016). Therefore, including each set of parameters improved model fit.

1) DIC Change for including/excluding random intercepts are not available, as model estimation without random effects capturing interindividual differences in these parameters failed.

No random effect for list was included because there is no meaningful interpretation of this effect (see explanation in Table S4).

2) Including the random effect for position (inter-individual differences in general performance increase or declines, e.g., by fatigue) in addition to the random intercept parameters (differences in recall performance) and random effect for the list x position interaction (inter-individual differences in Hebb learning) led to estimation problems for visuospatial material, therefore these random effects (variance plus related covariances) were excluded. In comparison to the GLMMs in Table S4 only half as many trials per participant (32 or 16 instead of 64 or 32 trials) were available so that the data only supported fewer random effects at the participant level. For verbal material including the random effect for position was possible, probably because each trial included 5 to 6 instead of 3 to 4 items. However, estimates of the random effects (see for example random effect variances for intercept 2 or List x Position) were already less stable, depending on whether the random effects for position were included or not. Nevertheless, estimates for the fixed effects (that are relevant for interpreting the results of the paper) were robust.

# **Section 4: Alternative analyses based on Normalized Levenshtein Distance**

During the review process it was pointed out that (normalized) Levenshtein Distance (see e.g., Kalm & Norris, 2016) is a particularly appropriate measure for operationalizing whether children improved in remembering the correct item sequence in Hebb trials as compared to filler trials. Given the reasons laid out in the method section and given that we preregistered that we would analyse the combined item and item in correct position recall scores, we ran all main analyses of the paper with these ordinal scores (0 = item not remembered; 1 = item remembered in any position; 2 = item remembered in position) as dependent variables and the preregistered cumulative logit models.

The mean scores of a trial giving credit for item and position information were highly correlated to normalized Levenshtein Distance scores (*r* = .94\*\*\*). To determine whether the results reported in the paper (and here in the supplement in Table S4/5 and S5/6) are robust and not only driven by item memory, we set up equivalent GLMMs for the Levenshtein Distance scores. As the normalized Levenshtein Distances have a restricted range from 0 to 1 like proportion correct scores, we disaggregated the scores from the trial to the item level and ran binomial GLMMs. For example, remembering the sequence A B C D E F as A C D B E F needs two edits to obtain the correct sequence, therefore LD = 2, and nLD = 1 - (2 / 6) = 0.67. The disaggregated values for that trial are 1 1 1 1 0 0 (the order of 1 and 0 within a trial is irrelevant for the analyses).

Table S6 and S7 display the results of these GLMMs on normalized Levenshtein Distance. Instead of two intercepts separating no recall vs. item recall in any position and item recall in any position vs. recall in position they contain only one intercept, leading also to fewer random effects that have to be estimated. As can be seen the fixed effects of these models on Levenshtein Distance scores and the models on item + position scores were all comparable (see comparisons within rows for the same sample / material). Of particular interest were fixed effects for the List x Position interaction representing the average Hebb learning effect. Irrespective of the scoring method, the parameter estimates were the same for the first two decimal places in the analyses run by group: ID: β = 0.05, *p*< .001; TD: β = 0.09, *p*< .001. The results were similar for the analyses run by material: verbal β = 0.09/0.10, *p*< .001/< .001; visuospatial: β = 0.03/0.03, *p*= .003/< .001 for the models based on normalized Levenshtein distance and on combined item + position scores, respectively.

Table S6. GLMMs on order recall measured by normalized Levenshtein Distance across trial positions for Hebb vs. filler lists to address research question 1 for the ID and TD groups (equivalent analyses to Table S4)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Normalized Levenshtein Distance Score** | | | | | | |  | **Ordinal Item + Position Score** | |
|  | **ID** | | |  | **TD** | | |  | **ID** | **TD** |
| Fixed Part | **β** | **95% CI** | ***p*** |  | **β** | **95% CI** | ***p*** |  | **β** | **β** |
| Intercept | -0.40 | [-0.62, -0.18] | < .001 |  | -0.42 | [-0.67, -0.17] | .001 |  |  |  |
| Material (verbal = +1 vs. visuospatial = -1) | 0.37 | [0.29, 0.45] | < .001 |  | 0.37 | [0.26, 0.49] | < .001 |  | 0.41 | 0.42 |
| List (Hebb = +1 vs. Filler = -1) | 0.18 | [0.11, 0.25] | < .001 |  | 0.06 | [-0.01, 0.13] | .08 |  | 0.16 | 0.03 |
| Trial Position (1st = 0, 2nd = 1 …) | -0.01 | [-0.04, 0.01] | 0.29 |  | -0.02 | [-0.05, 0.00] | .10 |  | 0.00 | -0.01 |
| List x Position (Hebb effect) | 0.05 | [0.02, 0.07] | < .001 |  | 0.09 | [0.06, 0.11] | < .001 |  | 0.05 | 0.09 |
| Random Part | **u** | **95% CI** |  |  | **u** | **95% CI** |  |  |  |  |
| *Random effect variances* |  |  |  |  |  |  |  |  |  |  |
| Intercept | 0.526 | [0.337, 0.820] |  |  | 0.617 | [0.389, 0.976] |  |  |  |  |
| Material | 0.059 | [0.033, 0.098] |  |  | 0.142 | [0.087, 0.226] |  |  |  |  |
| Trial Position | 0.003 | [0.001, 0.006] |  |  | 0.005 | [0.002, 0.009] |  |  |  |  |
| List x Position | 0.004 | [0.002, 0.006] |  |  | 0.004 | [0.002, 0.007] |  |  |  |  |
| *Covariances* |  |  |  |  |  |  |  |  |  |  |
| Intercept – Material | 0.007 | [-0.056, 0.070] |  |  | -0.041 | [-0.145, 0.056] |  |  |  |  |
| Intercept – Position | 0.015 | [-0.002, 0.033] |  |  | 0.017 | [-0.004, 0.040] |  |  |  |  |
| Intercept – List x Position | 0.037 | [0.022, 0.059] |  |  | 0.041 | [0.024, 0.067] |  |  |  |  |
| Material – Position | 0.000 | [-0.006, 0.007] |  |  | 0.008 | [-0.002, 0.02] |  |  |  |  |
| Material – List x Position | 0.001 | [-0.004, 0.007] |  |  | 0.000 | [-0.008, 0.009] |  |  |  |  |
| Position – List x Position | 0.002 | [0.000, 0.004] |  |  | 0.003 | [0.001, 0.005] |  |  |  |  |

Note. Sample sizes for the ID group: *N*Participants = 47, *N*Trials = 2944, *N*Items= 13680; for the TD group: *N*Participants = 47, *N*Trials = 2800, *N*Items = 13024

Table S7. GLMMs on order recall measured by normalized Levenshtein Distance across trial positions for Hebb vs. filler list for verbal and visuospatial stimuli to address hypothesis 3 (equivalent analyses to Table S5)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Verbal** | | |  | **Visuospatial** | | |  | **Ordinal Item + Position Score** | |
| Fixed Part | **β** | **95% CI** | **p** |  | **β** | **95% CI** | **p** |  | **Verbal: β** | **Visual: β** |
| Intercept | 0.02 | [-0.16, 0.19] | .84 |  | -0.88 | [-1.06, -0.71] | < .001 |  |  |  |
| List (Hebb = +1 vs. Filler = -1) | 0.12 | [0.06, 0.18] | < .001 |  | 0.13 | [0.05, 0.20] | .002 |  | 0.10 | 0.09 |
| Trial Position (1st = 0, 2nd = 1 …) | -0.03 | [-0.05, -0.01] | .006 |  | 0.01 | [-0.01, 0.04] | .26 |  | -0.03 | 0.02 |
| List x Position (Hebb effect) | 0.09 | [0.07, 0.11] | < .001 |  | 0.03 | [0.01, 0.06] | .003 |  | 0.10 | 0.03 |
| Random Part | **u** | **95% CI** |  |  | **u** | **95% CI** |  |  |  |  |
| *Random effect variances* |  |  |  |  |  |  |  |  |  |  |
| Intercept | 0.689 | [0.487, 0.960] |  |  | 0.579 | [0.398, 0.821] |  |  |  |  |
| Trial Position | 0.007 | [0.004, 0.011] |  |  | 0.004 | [0.001, 0.008] |  |  |  |  |
| List x Position | 0.007 | [0.005, 0.010] |  |  | 0.003 | [0.002, 0.005] |  |  |  |  |
| *Covariances* |  |  |  |  |  |  |  |  |  |  |
| Intercept – Position | 0.008 | [-0.014, 0.029] |  |  | 0.019 | [0.000, 0.037] |  |  |  |  |
| Intercept – List x Position | 0.052 | [0.034, 0.075] |  |  | 0.039 | [0.025, 0.056] |  |  |  |  |
| Position – List x Position | 0.002 | [0.000, 0.005] |  |  | 0.002 | [0.000, 0.004] |  |  |  |  |

Note. Sample sizes for the verbal condition: *N*Participants = 94, *N*Trials = 2880, *N*Items= 15792; for the visuospatial condition: *N*Participants = 94, *N*Trials = 2864, *N*Items = 10912

1. Model estimation including a pure position effect for Hebb trials succeeded in models on both sessions and groups, on both sessions in the ID group and on session 2 in both groups. Model estimation including a pure position effect only for Hebb trials failed in the model on both sessions in the TD group and on session 1 in both groups. As no residuals for session 1 could be extracted, retest reliability cannot be computed. [↑](#footnote-ref-1)